Administration

- HW due February 11 by 1 pm
- No class on Thursday, please bring HW to SS 2105
- Chapter 3: §3.1, 3.2 (except 3.2.4), 3.3 (except 3.3.3), 3.4 (except 3.4.4), 3.5.1
- Chapter 4: §4.1, 4.2, 4.3 (except 4.3.1, 4.3.2), 4.4.0, 4.4.1, 4.4.2
- Chapter 5: §5.1, 5.2, 5.3, 5.4, 5.5, 5.7, 5.9.0
- NR office hours are Tuesday 3-4 and Thursday 2-3
- BUT, Tuesday, will be late (SGS Exam) but will stay until 5; Thursday cancelled this week

Regression splines

- linear or generalized linear regression with derived feature variables
- allows responses to vary "smoothly" with features, without constraining (very much) "smooth"
- usual choice is to use cubic polynomials in windows of feature space, joining these continuously
- with linear fits at the ends of the range of the data
- fitted function (p. 146)

$$\hat{f}_j(X_j) = h_j(X_j)^T \hat{\theta}_j$$

- note change in notation from §5.2.1

... regression splines

- text p.146:

$$\hat{f}_j(X_j) = h_j(X_j)^T \hat{\theta}_j$$

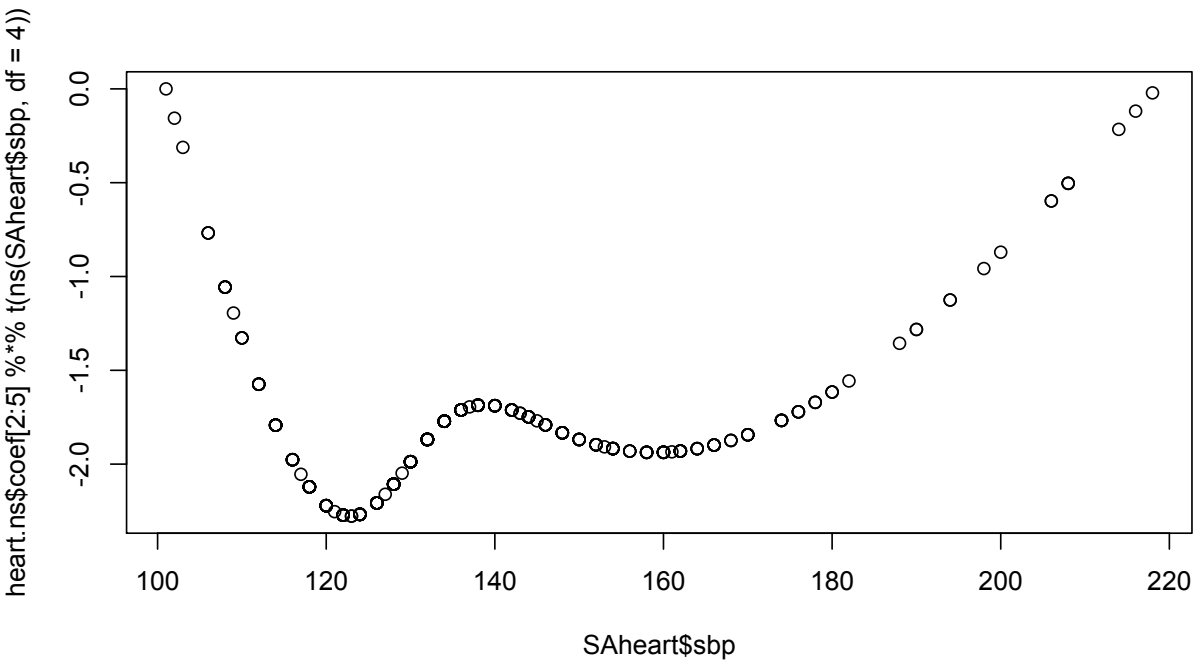
- previous notation (eqn. 5.2)

$$\hat{f}_j(X_j) = \sum_{m=1}^{M_j} \hat{\beta}_{jm} h_{jm}(X_j)$$

- in heart data example, 5 different fitted functions
  sbp, age, ldl, obesity, tobacco
- $M_j \equiv 4$; four derived variables for each feature
- in bone density example, a single covariate (age); $M_1 = 12$

Figure 5.4

```
> plot(SAheart$sbp, heart.ns$coef[2:5]%*%
+ t(ns(SAheart$sbp,df=4))
```



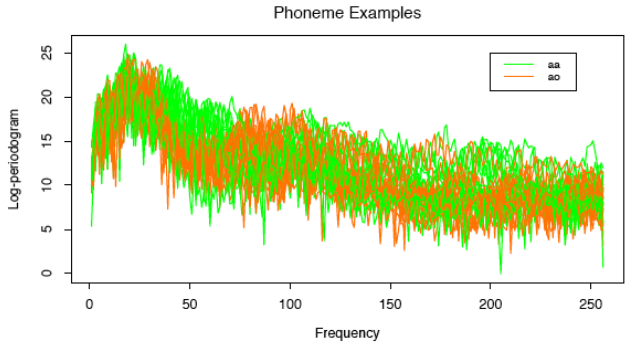
$$\hat{f}(sbp) = h_j(sbp)^T \cdot \hat{\theta}_j$$

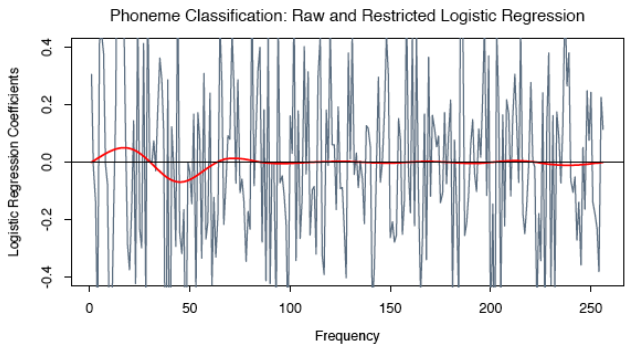
... Figure 5.4

```
> sefhatsbp = rep(0,462)
> for(j in 1:462){sefhatsbp[j] = sqrt(model.matrix(heart.ns)[j,2:5]
+ %*%vcov(heart.ns)[2:5,2:5]%*%model.matrix(heart.ns)[j,2:5]})
## doesn't reproduce Figure 5.4


Nancy

Here are the commands I used to produce that plot (essentially); the actual
plots were prettied up but otherwise they should be the same.
Trevor

postscript("nsglm2.ps",width=10,height=12,pointsize=14,horizontal=F)
par(mfrow=c(3,2),mar=c(5,5,4.2,1))
par(cex=.7)
fit.nsglm <- glm(chd ~ ns(sbp, 4) + ns(tobacco, 4) + ns(ldl, 4) + famhist +
ns(obesity, 4) + ns(alcohol, 4) + ns(age, 4), family = binomial, data =
heart[, c(1:3, 5, 7:9, 10)])
step.nsglm <- step(fit.nsglm)
plot.gam(step.nsglm,se=T,scale=8)# This scale just puts all the y axes on
the same scale
detach(2) ## plot.gam is in the library gam
dev.off()
```

Can you get this from the description on p.146? 

Regression splines as filters §5.2.3





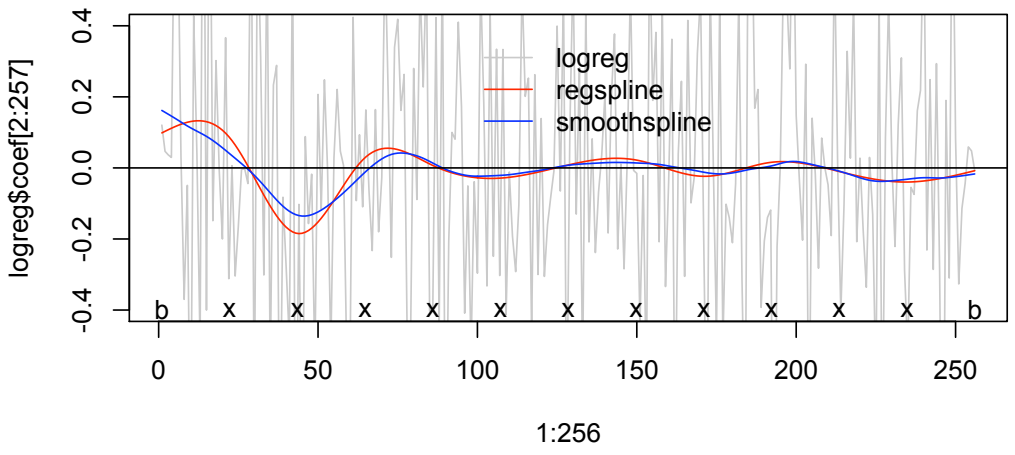

```
> phoneme[1,1:10]  # try phoneme[1,]
> phoneme[1, 250:258]  # we'll ignore the 'speaker' variable
> logreg = glm ( g ~ ., data = train[,1:257], family=binomial)
> smooth.fit = lm(logreg$coef[2:257] ~ ns(1:256, 12))
 ...
 see handout
```

Smoothing splines

- use natural cubic splines with knots at every observation
- penalize the coefficients
-

$$\hat{\mathbf{f}} = \{\hat{f}(x_1), \ldots, \hat{f}(x_N)\}^T = N(N^T N + \lambda \Omega_N)^{-1} N^T \mathbf{y} = S_\lambda \mathbf{y}$$

```
> library(MASS)
> smooth.fit2 = smooth.spline(1:256,logreg$coef[2:257],df = 12
> lines(smooth.fit2, col="blue")
```

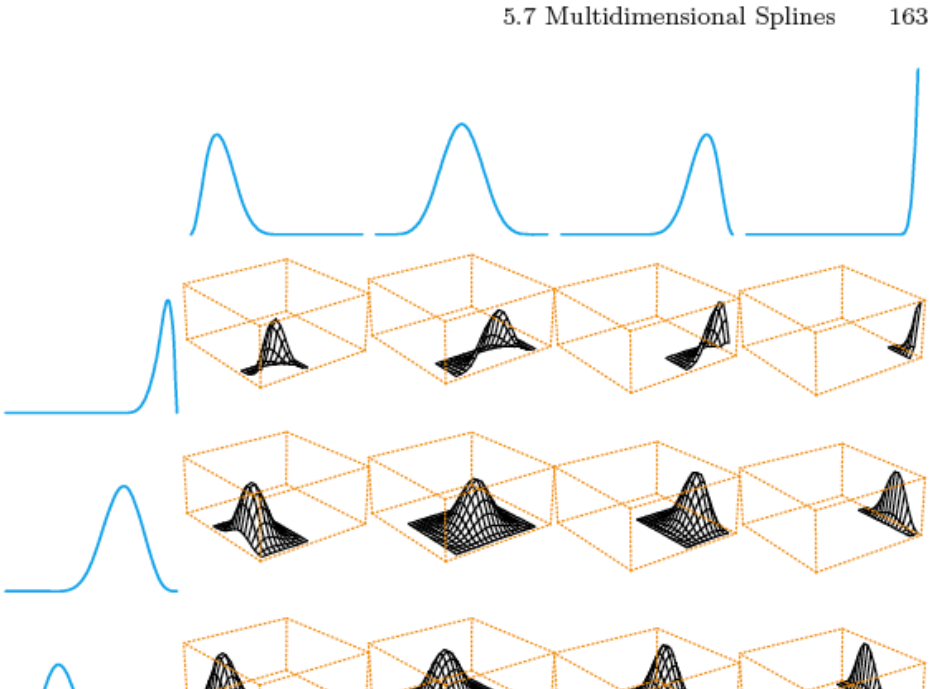


Multidimensional splines (§5.7)

- so far we are considering just 1 X at a time
- for regression splines we replace each X by the new columns of the basis matrix
- for smoothing splines we get a univariate regression
- with several X's we used additive models
- $E(Y \mid X_1, \ldots, X_p) = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$
- binary response:
  $\text{logit}\{E(Y \mid X_1, \ldots, X_p)\} = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$
  generalized additive models
- doesn't allow for interactions

... multidimensional splines

- regression splines with a two-dimensional basis can be constructed
- for example with all possible cross products: called tensor products
- $f(X_1, X_2) = \Sigma_{j=1}^{M_1} \Sigma_{k=1}^{M_2} \theta_{jk} h_{1j}(X_1) h_{2k}(X_2)$
- analogous to forming quadratic functions in regression using, e.g., $x_1^2, x_1 x_2, x_2^2$



... multidimensional splines

- smoothing splines in 2 dimensions

$$\min_f \Sigma_{i=1}^N \{y_i - f(\underline{x}_i)\}^2 + \lambda J(|f|)$$

- $J(|f|) = \int\int (\partial_1^2 f + \partial_2^2 f + 2\partial_{12} f)^2 dxdy$
- as in univariate case, solution exists in a spline basis similar to natural splines
- (5.39): $f(\underline{x}) = \beta_0 + \beta^T \underline{x} + \Sigma_{j=1}^N \alpha_j h_j(\underline{x})$
- $h_j(\underline{x}) = \eta(||\underline{x} - \underline{x}_j||), \quad \eta(z) = z^2 \log z$
- called radial basis functions: take this form because of symmetry of penalty
- uses N knots; reduced in implementation by regularization
- thin plate splines (Fig. 5.12)
- `library mgcv:  > help("mgcv-package")`

Smoothing splines vs regression splines: example

- The NMMAPS study Peng R., Dominici F., Louis T., (2006) JRSS A, 169, 179-203
- 90 largest cities in US by population (US Census)
- daily mortality counts from National Center for Health Statistics 1987–1994
- hourly temperature and dewpoint data from National Climatic data Center
- data on pollutants PM_{10}, O_3, CO, SO_2, NO_2 from EPA
- *output*: Y_t number of deaths on day t
- *inputs*: X_t pollution on day $t-1$, plus various confounders: age and size of population, weather, day of the week, time
- a model was fit for each city, and aggregated over cities
- **Conclusion** 0.41% increase in mortality for a 10 μg increase in PM_{10}

06/05/96 THE NEW YORK TIMES NATI

Data Revised on Soot in Air and Deaths

Scientists Lower Their Estimate of Risk From Bad-Air Days

By ANDREW C. REVKIN

Revisiting their own data with new methods, scientists who conducted influential studies that linked sooty air pollution with higher death rates have lowered their estimate of the risk posed by bad-air days.

The findings do not challenge what is now a well-established link between air pollution and premature death. But the new analysis is highly likely to delay the final review of new regulations on small-particle pollution, officials of the Environmental Protection Agency said yesterday.

The review was projected to end, and the new rules to take effect, by the end of next year.

"This may clearly push it beyond that," a spokesman for the E.P.A., Joe Martyak, said last night.

The fine particles in question come mainly from power plants and diesel engines, and the proposed rules have been at the center of a long legal, political and public-relations battle between private environmental groups and power plant owner and vehicle manufacturers.

The researchers, at the Johns Hopkins University, have been distributing their new analysis to scientists and government officials by fax and e-mail. Yesterday, they set up a Web site, biosun01.biostat.jhsph.edu/~fdominic/research.html, that details their new findings.

cles that can be deeply inhaled into the lungs and stay there. In the original analysis, the rise was 0.4 percent above the typical mortality rate for each jump of 10 micrograms of soot per cubic meter of air. In the new analysis, the increase is half that.

The researchers said the change was small but significant. The average level in the average city is now about 24 micrograms a cubic meter.

The work has been published for several years in a variety of leading journals like The New England Journal of Medicine and The American Journal of Epidemiology. The project, the National Morbidity, Mortality and Air Pollution Study, was given extra weight by policy makers because of the reputation of the Health Effects Institute and the Johns Hopkins group, led by Dr. Jon-

New research may delay a review due next year on small-particle pollution.

athan M. Samet, chairman of epidemiology at the public health school

kins biostatistics department, Dr. Scott L. Zeger, said other researchers who used the software, S-Plus, should check for similar problems. It is widely used for research in fields like pharmacology, genetics, molecular biology and stock-market forecasting, as well as serving as a mainstay of other environmental studies.

Dr. Zeger and Mr. Greenbaum stressed that the new findings did not overturn the basic link between soot and illness. They also pointed to the recent publication of other studies on the long-term effects of soot that do not use the same analytical tools.

Still, industry officials said the new findings called into question the validity of some research underlying the new federal standards.

"This study is really one of the ones creating the path for the future on air-quality regulation," said Allen Schaeffer, executive director of the Diesel Technology Forum.

The new results, Mr. Schaeffer said, show that "particle science is still evolving, and so are the analytical tools to look at it."

Scientists involved with the soot standard said that there was much other evidence that pointed to the dangers of soot but that the errors in the Johns Hopkins work were still significant.

"It certainly brings into question the precision of the data," said Dr. Jane Q. Koenig, a professor of envi-

Statistical error leaves pollution data up in the air

Jonathan Knight, San Francisco

An off-the-shelf statistics package has tripped up pollution researchers in North America and Europe who are studying the effects of airborne soot on human health.

A default setting that produced erroneous results went unchecked for years, despite significant statistical expertise in all of the groups. "It was already such standard software when we started using it, I didn't even question it," says Francesca Dominici, a public-health researcher at Johns Hopkins University in Baltimore, Maryland.

Clouding the issue: particulates are seen as a health risk, but how reliable are the data?

On 4 June, Dominici posted revised figures on her website after discovering that the error had doubled her group's estimate of the risk to health posed by particulates in the air. Two other groups that used the same tool, one in Canada and one in Greece, are now redoing their calculations.

The groups were looking for correlations between death rates and particulates in the air, which come mainly from diesel engines and power plants. Their data on air quality, hospitalizations and deaths from dozens of cities cover a seven-year period up to 1994.

Death rates vary throughout the year because of such factors as changes in temperature and disease outbreaks. To tease out the effects of particulates, the groups used a statistics program known as S-Plus.

S-Plus searches for correlations using an iterative process in which confounding effects are gradually peeled away. The default parameter in question determined how many times the procedure would iterate before stopping to produce a final result.

"For most applications the value is perfectly fine," says David Smith, product manager of Seattle-based Insightful, which sells S-Plus. Smith says that the Hopkins case was exceptional, but that users should always check whether changing the parameter affects the outcome, and adjust it if necessary. Smith says that Insightful will tighten the default value of the parameter — slowing the programme slightly — on future releases of S-Plus.

Richard Burnett, a statistician with Health Canada in Ottawa, which is conducting a similar study, says that his group will probably revise its estimates of the impact of airborne soot on mortality downwards by 20–50%. The findings of a study run by a group at the University of Athens may also have to be adjusted, he says.

The health risk posed by particulates is a source of fierce environmental controversy in the United States, and the Bush administration is considering rules to restrict emissions. Opponents of tighter rules are likely to seize on the revisions as evidence that the research linking soot in the air to harmful effects on health is not to be trusted. ■

... the model

- $Y_t \sim Poisson(\mu_t)$
- $\log \mu_t = \beta X_{t-1} + \gamma DOW + s_1(t, 7) + s_2(temp_0, 6) + s_3(temp_{1-3}, 6) + s_4(dew_0, 3) + s_5(dew_{1-3}, 3)$
- $s(x, 7)$ a smoothing spline of variable x with 7 degrees of freedom (`gam`)
- estimate of β for each city; estimates pooled using Bayesian arguments for an overall estimate
- very difficult to separate out weather and pollution effects
- relevant: Crainiceanu, C., Dominici, F. and Parmigiani , G. (2006). Adjustment Uncertainty in Effect Estimation `http://www.biostat.jhsph.edu/~fdominic/research.html`
- mortality rates change with season, weather, changes in health status, ...
- problem with convergence criterion and standard errors
- new estimates used regression splines and `glm`

- “the new analysis is highly likely to delay the final review of new regulations on small-particle pollution”
- “industry officials said the new findings called into question the validity of some research underlying the new federal standards”
- “ ‘It certainly brings into question the precision of the data’, said Dr. Jane Q. Koenig”
- “The health risk posed by particulates is a source of fierce environmental controversy in the United States”
- “Opponents of tighter rules are likely to seize on the revisions as evidence that the research linking soot in the air to harmful effects on health is not to be trusted”
- “A default setting that produced erroneous results went unchecked for years, despite significant statistical expertise in all the groups”

- "The findings do not challenge what is now a well established link between air pollution and premature death"
- "The work has been published for several years in a variety of the leading journals like the New England Journal of Medicine and the American Journal of Epidemiology"
- "The project, the National Morbidity, Mortality and Air Pollution Study, was given extra weight by policy makers because of the reputation of the Health Effects Institute and the Johns Hopkins group"
- **Not as well known that the problem was first discovered at Health Canada, by Tim Ramsay and Rick Burnett**
- their work also drew attention to the incorrect calculation of standard errors in the `gam` software
- Original estimate 0.41% increase in mortality rate associated with increase of $10\mu g/m^3$ increase in PM_{10}.
- Revised estimate 0.22%.

- these are small effects; approximately 15 additional deaths per year in Toronto
- current software in `R` using `library mgcv` has solved the problem with smoothing splines
- most current work on pollution effects now uses regression splines
- Figures from revised NMMAPS study: nmmaps-revised.pdf
- `library(NMMAPS)` etc.

Wavelet bases

- regression spline basis uses set of smooth functions
- e.g. `matplot(SAheart`*sbp*, *ns*(*SAheart*`sbp,4)`
- smoothing spline basis similar, but larger (knots at each distinct x value)
- polynomial or orthogonal polynomial basis can also be used, but non-local
- Fourier basis often used in time series
- wavelet basis composed of highly localized pieces Figure 5.16
- basis functions are orthogonal
- fit a model with many basis functions, then throw away all those with small coefficients: thresholding
- handout